Journal of Quantitative Description: Digital Media 2(2022), 1–35 10.51685/jqd.2022.002

A Note on Increases in Inattentive Online Survey-Takers Since 2020

JOHN TERNOVSKI
LILLA ORR
Yale University, USA

We present empirical evidence showing that, since 2020, there has been a significant increase in inattentive responses on Lucid, a popular source of online convenience survey samples. Inattentive participants – respondents who incorrectly answer directed query attention check questions – may be introducing substantial measurement error and attenuation bias. Using data from 152,967 survey respondents across multiple studies conducted between January 2020 and June 2021, we find that inattentive respondents report less reliable demographic data, less stable responses, and are systematically different from attentive respondents. We find some evidence of attenuation bias and mixed evidence that data quality has decreased slightly since 2020 even after excluding inattentive respondents. We conclude that researchers using Lucid should report if they screened on attentiveness and consider replicating any null results. Such an unexpected increase in inattentiveness in a widely-used platform suggests that future researchers relying on online convenience survey samples should continually assess data quality.

Keywords: data quality, attentiveness, attention checks, crowdsourcing, online surveys

John Ternovski: john.ternovski@yale.edu
Date submitted: 2021-07-21

The social sciences have long used individual subjects' self-reported sentiments, beliefs, and attitudes for empirical research. At the heart of this research has always been the question: to what extent do responses to survey questions accurately describe what people actually think, feel, and/or believe? The integrity of the mapping of a person's internal state to survey data has been described as the amount of error in the data, or, conversely, the quality of the survey data (e.g., Grove, 1987). Decades of work by survey methodologists has produced polling best-practices (e.g., question design, sample selection, etc.) to maximize survey data quality (for an overview, see Atkeson & Alvarez, 2018). But since traditional, nationally-representative polls have tended to be expensive, and social scientists often set out to study "ubiquitous and universal" phenomena, convenience samples of undergraduate students became a common, inexpensive alternative for many academic researchers (Sears, 1986; p. 519). In the 2000s, however, the rise of online survey platforms like Amazon Mechanical Turk (MTurk) promised an even more convenient, more inexpensive sample that was not necessarily limited to 18-22-year-olds (Buhrmester, Kwang & Gosling, 2011; Berinsky, Huber, & Lenz, 2012).

The use of online survey platforms by academics (sometimes referred to as "crowdsourcing" research) has rapidly become highly prevalent in the social sciences (Mullinix et al., 2015). Analyses of studies published in top psychology journals between 2015 and 2018 found that around 50% of those studies made use of online convenience samples (Anderson et al., 2019; Zhou & Fishbach, 2016; Buhrmester, Talaifar, & Gosling, 2018). Other social science fields saw similarly high rates of crowdsourcing research (e.g., 43% of studies in consumer research between 2015 and 2016 (Goodman & Paolacci, 2017) and as much as 31% of studies in cognitive sciences journals in 2017 (Stewart, Chandler, Paolacci, 2017)).

Early evaluations of the quality of crowdsourced research yielded highly promising results across disciplines. Prominent, highly-cited papers in psychology (Buhrmester, Kwang & Gosling, 2011), political science (Berinsky, Huber, & Lenz, 2012; Mullinixet al., 2015; Levay, Freese, & Druckman, 2016; Coppock, 2019; Coppock, Leeper, & Mullinix, 2018), consumer research (Goodman & Paolacci, 2017), and medicine (Morensen & Hughes, 2018) validated and endorsed the use of crowdsourcing research in their respective fields. Crowdsourced data were found to be "more demographically diverse" than typical samples in psychology (i.e. predominantly undergraduate students) (Buhrmester, Kwang & Gosling, 2011) and political science (Berinsky, Huber, & Lenz, 2012). More importantly, crowdsourced survey experiments appeared to have similar treatment effects to that of nationally-representative survey experiments (Mullinix et al., 2015) and in-person labs (Amir, Rand & Gal, 2012).

These early evaluations primarily assessed the most prominent online survey platform for crowdsourced data, Amazon Mechanical Turk (MTurk). To ensure that the participants were not answering questions haphazardly, it became common practice to limit entry to only those MTurk participants who completed many studies in the past (over 100) and received overwhelmingly positive "reviews" (i.e., 95%+ approval rates) from other researchers.[1] Perhaps partially due to this norm, researchers later found that about 10% of all MTurk users are responsible for 40% of all survey responses (Chandler, Muller & Paolacci, 2014) and the total population from which the typical MTurk study samples was estimated to be just 7,300 individuals (Stewart et al., 2015). These experienced MTurk users are likely to be exposed to many different social science studies and may became familiar with the types of questions and randomized interventions researchers commonly use. Prior repeated exposure to a particular survey instrument may change future responses without the researcher ever knowing (Chandler, Muller & Paolacci, 2014). Indeed, Chandler et al. (2015) found that "[w]hen participants in the present study performed the same tasks on two different occasions, effect sizes decreased by about 25% at the second time point" (p. 1137). Additionally, recent research has indicated that MTurk participants may be engaging in "fraud." (Dennis, Goodson, and Pearson, 2020). Namely, Dennis,

[1] This norm has both been reported (e.g., Berinsky, Huber, & Lenz (2012) note that "[r]equesters often specify at least a 95% prior 'approval rate'" (p. 366)) and reinforced (e.g., Buhrmester, Talaifar, & Gosling (2018) recommend that researchers set approval rates to 95% or more).

Goodson, and Pearson (2020) find that MTurk participants can use technological loopholes (such as Virtual Private Servers [VPSs])[2] to provide duplicate data and misrepresent their identities (e.g., where they are from). These data quality issues on MTurk may have been one of the reasons alternative crowdsourcing platforms have since recently grown in popularity among social scientists.[3] This paper looks at one such alternative: Lucid Theorem.

Lucid is a source of online survey panels with participants recruited from a wide variety of platforms ranging from mobile games to market research survey websites. It is an increasingly popular alternative to MTurk. In a recent empirical validation, Coppock & McClellan (2019) concluded that "Lucid boasts a much larger pool of subjects than MTurk; the risk of cooperation among subjects is minimal given their diverse sources; subjects are less professionalized; subjects are more similar to US national benchmarks in terms of their demographic, political, and psychological profiles. Experimental results obtained on Lucid are solidly in line with the results obtained on other platforms." (p. 12). Use of Lucid grew rapidly and data sourced from Lucid has contributed to empirical research in a variety of fields including political science (e.g., Wood & Porter, 2019), psychology (e.g., Pennycook et al., 2020), economics (e.g., Benzell, Collis, & Nicolaides, 2020), public policy (e.g., Hemel & Porter, 2020), environmental science (e.g., Motta et al., 2019), and medicine (e.g., Solnick, Peyton, & Kraft-Todd, 2020). Lucid Theorem have recently begun compiling a list of academic studies that use Lucid data, which they present on their website. This very rough benchmark nevertheless gives some sense of Lucid's growing prominence in academic research: there was one study in 2018, three in 2019, 45 studies in 2020, and 22

[2] Since there is no universal unique identifier for individual people online, a commonly-used approximation is the IP address. An IP address denotes the network location from where a computer is accessing a given website; that network location is usually tied to geographic coordinates. However, there exist technological solutions that allow a user to adopt an IP address associated with a different geographic location. The most common such solutions are Virtual Private Servers (VPSs) and Virtual Private Networks (VPNs). Both services allow a user to temporarily replace their device's IP address with an IP address that is associated with a computer at a data center elsewhere in the world. (The main relevant difference between VPN and VPS usage is that VPN usage generally involves multiples users sharing the same IP address.)

[3] Kennedy et al. (2020) propose matching IP addresses to public databases of VPSs and add-on service vendors like TurkPrime have begun offering this kind of IP-based screening service for a fee.

studies in the first four months of 2021 (Lucid, 2021). Sixteen of these studies have been published in top political science journals (6 in the American Journal of Political Science, 3 in American Political Science Review, and 7 in Journal of Politics).[4] But much like MTurk required repeated and ongoing validation of data quality as it grew in popularity, here, we reassess the quality of data from Lucid in light of its growth and the COVID-19 pandemic.

In this paper, we find that, across studies conducted by multiple independent research teams, with 152,969 consenting respondents in total, the quality of the data on Lucid appears to have dropped significantly in 2020. We find that rates of inattentiveness rose in the first six months of 2020, and that respondents who fail an attention check provide less reliable demographic data, less stable responses across identical questions, and are systematically different from respondents who pass. We further find that an experiment replicated with attention screening found significantly increased treatment effects compared to a sample that did not screen for inattention, suggesting that inattention on Lucid may lead to attenuation bias. We also find mixed evidence that data quality may have declined slightly even among participants who successfully pass attention checks. Lastly, we find no evidence that data quality has meaningfully bounced back to early-2020 levels in the first half of 2021.

Since our research suggests that the quality of online data may vary over time even on popular, previously-validated platforms, we recommend that researchers take measures to detect and correct for low quality data. Low quality data may introduce sizable measurement error to survey outcomes and may attenuate treatment effects. Specifically, we recommend researchers conducting surveys on Lucid (or any other online survey platform) consider incorporating attention checks and assessing the quality of their data (such as by asking participants' age or location to match against the demographic variables on file with Lucid). Researchers who have conducted studies on Lucid without attention

[4] All but one of these studies were published in 2020 or later. A full list of the studies (including dates) is available in the Appendix.

checks should consider replicating any null results after screening on attentiveness. At a minimum, researchers should transparently report whether or not they use attention checks or similar screens for data quality in any publications.

Assessing Data Quality

We consider three major challenges associated with using data from online survey platforms: 1) uniqueness/independence (i.e., survey responses are unique to individuals in the sample and individuals do not influence each other’s responses), 2) identity (i.e., participants are who we think they are), and 3) legitimacy of responses (i.e., respondents read the question and answer to the best of their ability).

Uniqueness/Independence

First, it is difficult to establish that any given participant is a unique, independent subject. Since there is a financial incentive to participate in an online survey, a malicious actor may attempt to disguise oneself online and take the same survey multiple times (see Dennis, Goodson, and Pearson (2020) for a more extensive discussion).[5] And if this phenomenon is widespread, it could have serious implications for online survey experiments, since subjects may inadvertently be exposed to multiple experimental treatments.

Lucid appears to have taken steps to prevent this. When we tried to take Lucid surveys through a Virtual Private Network (VPN), a software tool that effectively disguises the online identity of the participant, we found that Lucid did not allow us to participate in

[5] A related challenge to independence is that participants may communicate with one another and one participant in a survey may share information with another participant in the survey. For instance, there are forums of MTurk users, where studies are discussed by participants while those studies are running. This issue is more likely to be a problem with MTurk, due to the professionalization of users and the relatively small population from which samples are drawn (Dennis, Goodson, and Pearson, 2020).

any surveys.[6] While this does not necessarily ensure that there are no duplicate participants in a study, it does provide a platform-level safeguard. To further protect against respondent duplicates, researchers may also use the tools compiled in Kennedy et al. (2020) to match their survey participants' IP addresses to a public database that has information about the likelihood that the IP address is a VPS or proxy.

Identity

It is usually important to have some general information about who is taking the survey. For instance, a survey experiment on the political attitudes of Americans would yield misleading conclusions if a majority of the sample were, in fact, non-American (see Dennis, Goodson, and Pearson (2020) for more on online identity verification). So, a second challenge is ensuring that participants are who they say they are. This is particularly an issue for studies that rely on demographics or other individual characteristics (e.g., using demographics to weight a sample or assessing heterogenous treatment effects).

Without de-anonymizing participants, the researcher has two main ways of gathering evidence that the participant is who they say they are: 1) confirming demographic information collected by the online survey platform against self-reported information collected during the course of the survey and 2) checking for consistency in self-reported information collected during the survey.

The first approach uses data that has been collected by the online survey platform before the participant is able to participate in surveys. For instance, Lucid gathers basic demographics about all respondents and provides them to the researcher. Asking participants to self-report any of the fields that Lucid stores can provide some evidence that the identity is, at least, consistent. Relatedly, a researcher can ask participants the same question multiple times, potentially with different question wordings. For instance, a survey can ask for birth date and age; a mismatch may indicate that the age provided may

[6] It appears that Lucid uses a blacklist of known VPN IP addresses.

not be accurate (though, in some cases, it may also simply be a good-faith error by the participant, or participants may misreport personal information in the same way each time they are asked).

Legitimacy of Responses

The third challenge is ensuring that the responses are "legitimate" answers to survey questions. This means that the participant read the question and honestly responded to the best of their ability. However, this goal of the researcher may be incompatible with the incentives of the participant. As discussed in detail in the canonical work on satisficing (Krosnick & Alwin, 1987), a participant in survey research has strong incentives to put minimal cognitive effort into satisfactorily completing a survey. An extreme example of this sort of behavior would be selecting answers haphazardly without reading the questions. This could introduce meaningfully high levels of measurement error and, in randomized survey experiments, could potentially obfuscate any treatment effects through attenuation bias. We should emphasize that this issue is not unique to crowdsourced research and there is ample research on this phenomenon across different modes of interview (for a review, see Roberts et al. (2019)).

There have generally been three major approaches to detecting satisficing behaviors in surveys: 1) consistency across similar questions, 2) speed at which a respondent completes the survey, and 3) incorporating "attention check" questions throughout the survey.[7]

The first strategy is similar to one of the strategies researchers can use to confirm an individual participant's identity. If an individual gives very different answers to two very similar questions, this may indicate that they are not reading the question and/or

[7] There also exist group-level and study-level approaches to detecting satisficing (e.g., replicating canonical treatment effects (Coppock & McClellan, 2019)), but since these approaches cannot identify satisficing individuals, they are not as useful for researchers who are trying to screen and exclude satisficing respondents before they complete the survey. As such, we refrain from discussing them here.

selecting answers haphazardly (e.g., Wood et al., 2017). For instance, asking how favorable a respondent feels towards Joe Biden and asking how much a respondent approves of the job Joe Biden is doing as president should yield highly correlated responses since both questions are designed to measure the underlying construct of an individual's opinion of Joe Biden. A low or negative correlation may indicate an issue in data quality.

The second strategy tracks how fast a participant completes a specific question or, more commonly, the entire survey (e.g., Malhotra, 2007; Read, Wolters & Berinsky, 2020). The intuition is that the participant who is satisficing is more likely to speed through a survey without reading the questions and considering their responses. The main issue with this (and the response consistency strategy) is that there has not been sufficient consensus in the literature to indicate the threshold at which a participant's responses should be discarded due to suspected satisficing. In other words, how fast must a participant complete a given survey for a researcher to lose confidence in the legitimacy of that participant's data? Furthermore, it is challenging to compare such thresholds across different surveys of varying levels of length, difficulty, complexity, and the population sampled.

As such, one of the most popular strategies to detecting satisficing is embedding "attention check" (or directed query) questions throughout the survey (e.g., Alvarez et al., 2019; Paas & Morren, 2018; Kung, Kwok, & Brown, 2018; Gummer, Roßmann, & Silber, 2018; Meade and Craig, 2012). The directed query is one of the most straightforward attention checks: to show that participants have read a question, the survey asks participants to select a particular response (e.g., "For this question only, select the choice Strongly Disagree. Do not select any other choice."). Any participant who fails the (pre-treatment)[8] attention checks can be excluded from the analysis.[9] To be explicit, failing an attention check does not necessarily mean that that participant provides only nonsensical data. Attention may wax and wane over the course of a survey and even generally diligent

[8] Screening on post-treatment attention checks can introduce bias that results from asymmetric samples between treatment arms (Aronow, Baron, & Pinson, 2019).

[9] For detailed recommendations on attention check usage, see Berinsky et al. (2021).

survey-takers could occasionally fail an attention check (e.g., Read, Wolters & Berinsky, 2020).[10] However, prior research has shown that, on average, attention check failure is associated with reduced data quality in terms of:

1. correlational consistency (Berinsky, Margolis & Sances, 2014; Alvarez et al., 2019; Gummer, Roßmann, & Silber, 2018),
2. recovering canonical (i.e., widely-replicated) treatment effects (e.g., Berinsky, Margolis & Sances, 2014; Peyton, Huber, & Coppock, 2021),
3. speeding through the survey (Alvarez et al., 2019; Paas & Morren, 2018; Gummer, Roßmann, & Silber, 2018),
4. "straight-lining" or answering the same multiple choice (e.g., "Strongly agree") for all questions (Paas & Morren, 2018; Gummer, Roßmann, & Silber, 2018),
5. and giving implausible answers (Gummer, Roßmann, & Silber, 2018).

Additionally, failing one attention check early on in a survey predicts failing a second attention check at the end of the same survey (Paas & Morren, 2018). In short, while screening on attention checks may result in false positives (i.e., excluding "good" data), there is compelling evidence to indicate that such screens exclude participants who do not read all questions and/or select answers haphazardly.

The attention check does not address all forms of satisficing, but rather allows researchers to better identify a subset of satisficing participants who are likely to answer other questions carelessly or haphazardly. This paper primarily addresses this form of satisficing and we refer to this subtype as inattentiveness. Though attention check failure is only an estimate of latent inattentiveness, for brevity, we refer to participants who fail attention checks as "inattentive participants" throughout this paper.

The use of attention checks is far from universal. Other scholars caution that excluding based on failed attention checks may exclude legitimate responses, reduce

[10] An attention check failure may also be the result of a mis-click.

power, and reduce the sample's representativeness (e.g., Downs, Holbrook, & Peel, 2012; Abbey & Meloy, 2017; Berinsky, Margolis & Sances, 2014); their very presence may even subtly affect outcomes (e.g., Hauser, Ellsworth, Gonzales, 2018; Hauser & Schwarz, 2015). While addressing these drawbacks is outside the scope of this research note, we maintain that the main advantages of filtering participants on pre-treatment attention checks should also be considered. The advantages are that this exclusion keeps survey costs low (i.e., on Lucid, researchers do not need to pay for respondents who fail attention checks) and reduces measurement error due to inattentiveness—how these advantages compare against potential disadvantages will depend on the specific context and research question. In the case of Lucid surveys since 2020, we believe that the magnitude of measurement error due to inattentiveness is high enough to warrant the routine use of attention check exclusions.

Data

The data in this paper are sourced from four independent online survey studies. All four are multi-wave, rolling cross-sectional studies. The relevant characteristics of each study are described in turn.

Kalla (Study K)

Kalla (in separate studies with David Broockman and Micah English) conducted surveys on Lucid from January 6, 2020 to June 23, 2021 with a total of 119,172 participants. Every wave of the survey began with the same three questions: a consent-to-participate question, a question asking participants to pay careful attention and mark "I understand", and a variant of a commonly used attention check:[11]

> People are very busy these days and many do not have time to follow what goes on in the government. <u>**We are testing whether people read**</u>

[11] All subsequent attention checks are presented verbatim and use the exact text formatting (i.e., bolding, underlining, etc.) actually used in each survey.

questions. To show that you've read this much, answer **both** "extremely interested" and "very interested."

Extremely interested
Very interested
Moderately interested
Slightly interested
Not interested at all

Ternovski (Study T)

Ternovski (with Joshua Kalla and P. Aronow) conducted 7 waves of surveys on Lucid between April 12, 2020 and November 8, 2020 (N=12,279). After participants consented and confirmed their eligibility status (e.g., 18+ years of age), there were two audiovisual checks.[12] Participants then had to pass two variants of standard attention checks:

1. For our research, careful attention to survey questions is critical! To show that you are paying attention please select "I have a question."

   I understand
   I do not understand
   I have a question

2. People are very busy these days and many do not have time to follow what goes on in the government. We are testing whether people read questions. To show that you've read this much, answer both "extremely interested" and "very interested."

[12] Participants were shown a test video with sound where an actor held up 8 fingers and said, “one hundred and twenty-three.” On the next page, participants were then asked how many fingers the actor held up and what was the number that was said aloud. If they answered either question incorrectly, they were given an opportunity to watch the video again and attempt the same questions on the next page. Individuals who gave an incorrect answer on the second attempt were screened from the survey. Since we are unable to differentiate whether failure at this juncture is due to technical difficulties or inattentiveness, we refer to these screens as audiovisual checks, rather than attention checks.

Extremely interested
Very interested
Moderately interested
Slightly interested
Not interested at all

Schaffner (Study S)

Schaffner conducted 12 waves of online surveys on Lucid from April 11, 2020 to July 28, 2020 (N=14,304). All waves included three attention checks. Unlike the above studies, a respondent was considered passing if at least two of these three attention checks were answered correctly. The questions were embedded in separate grids, and asked respondents to “Please just select [negative/positive]”, “Please just click [Oppose/Support]”, and “Please just select slightly decrease” with the direction varying across waves.

Orr (Study O)

Orr conducted two nearly identical studies on Lucid—one was fielded May 4-5 and May 22-28, 2020 and had no attention checks (N=2,200). A second version was fielded July 14-18, 2020 and included an attention check screen (N=5,014). To be consistent with the other studies, we refer to the two Orr studies as waves. In the second wave, respondents were only allowed to complete the study if they passed a similar attention check to that used in Study K and Study T:

> People are very busy these days and many do not have time to follow what goes on in the government. **We are testing whether people read questions.** To show that you’ve read this much, answer **both** “extremely interested” and “very interested.”
>
> Extremely interested

Very interested
Moderately interested
Slightly interested
Not interested at all

Results

The results are organized as follows. First, we show that the rate at which participants pass attention checks declined in 2020. We then make explicit why these declines can be consequential by estimating the levels of measurement error associated with inattentive responses and show that failing to use attention screens may attenuate treatment effects. We then illustrate that, consistent with prior research, inattentive participants are systematically different from their attentive counterparts across common demographic variables like age, sex, and socioeconomic status, but we note that inattentive participants may also be providing spurious demographic information. We then show that attention checks are not a panacea to poor data quality and that there is mixed evidence that measurement error may be increasing even for attentive participants. We conclude with a 2021 update, which suggests that data quality did not meaningfully improve in the first half of 2021, after the height of the COVID-19 pandemic.

Attention Check Success Rates Over Time

Three of the studies in this analysis had attention checks across multiple waves (K, T, and S) and all three studies exhibited declines in attention check success rates in the first 6 months of 2020 (Figure 1). The final waves of Study T showed an increase in success rates in October and November of 2020.[13]

[13] Studies S and K did not collect data in those months.

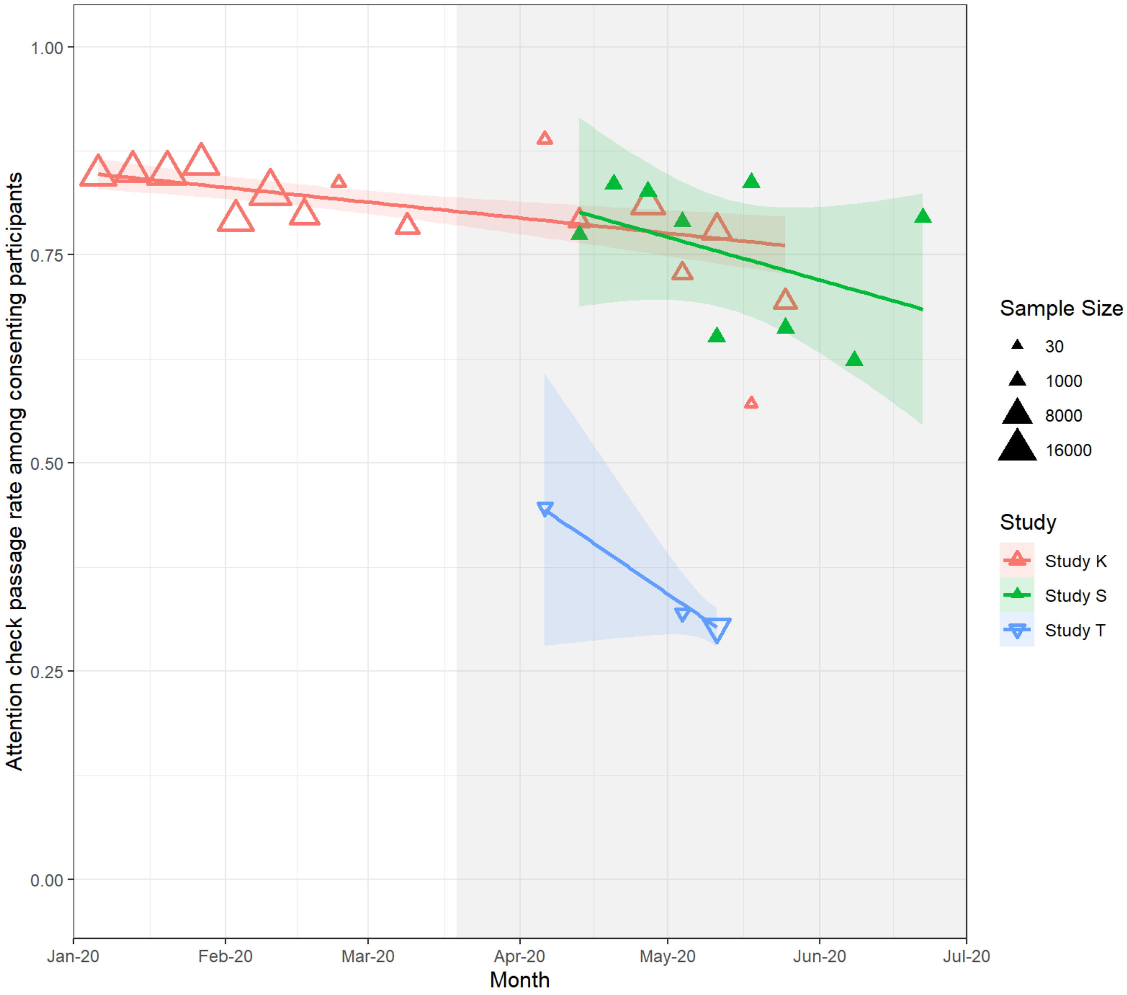

Figure 1. Attention Check Passage Rate Decreased Over the First 6 Months of 2020. *Note*. Figure 1 plots weekly attention check passage rates over the first 6 months of 2020 including any week in which more than 10 responses were recorded. In Study T, passage rates include attention check and audiovisual check passage. The time trend within each study is summarized through linear regression, weighted by weekly sample size and displayed with 95% confidence intervals. Study K β = -0.004 average weekly change in passage rate, p < 0.01, Study S β = -0.01, p = 0.31, Study T β = -0.03, p = 0.06. The shaded region indicates the period in which widespread precautionary measures were taken in the United States to combat the Covid-19 pandemic, beginning March 19th, 2020.[14]

[14] The uptake of precautionary measures was gradual. On March 19, California implemented a statewide lockdown, "mandating all residents to stay at home except to go to an essential job or shop for essential

It is important to note that none of these studies were panels (i.e., longitudinal studies), so each wave within a study was a new sample drawn from a population of Lucid survey-takers. The survey solicitations were identical in every wave, so it is unlikely that temporal trends within study can be attributed to differences in survey characteristics.[15] These declines in attentiveness may be due to a changing population of Lucid survey-takers (i.e., more inattentive participants to sample from) and/or the population of Lucid survey-takers grew more inattentive, on average (i.e., the composition of the population of Lucid survey-takers did not change but survey-takers became more inattentive).

As a rough benchmark, attention check success rates in other online survey contexts[16] (usually Mturk) have generally been reported to be around 80%-90% (e.g., Conn, Mo, & Seller, 2019; Hauser & Schwarz, 2016; Kung, Kwok, & Brown, 2018; Abbey & Meloy, 2017; Silber, Danner, & Rammstedt, 2019). The low success rates in Study T are noteworthy for digital media studies; even in waves with the highest rates of success, the majority of Lucid respondents were unable to confirm that they could see video, hear sound, and read text.

Inattentiveness and Measurement Error

This decline is problematic, as those who fail attention checks may introduce substantial measurement error. To evaluate measurement error, or the legitimacy or responses, one would ideally compare an observed survey response to the truth. Since the true value is usually unobservable, the best we can do is examine response consistency. Namely, we can examine whether participants' responses to the same questions change

needs" (AJMC Staff 2021, p. 1). Other states and local governments followed suit with a total of 42 states and territories issuing mandatory stay-at-home-orders before the end of May 2020. (Moreland et al. 2020, p. 1198).

[15] We urge caution when comparing attention success rates across the three studies. Differences in the wording of each study's solicitation, description, and consent forms may lead to sorting, which could explain some of the differences in attention success rates in the same window of time between studies.

[16] For questions similar to those used in Studies S and K.

over the course of the same survey (Cf. Ansolabehere, Rodden, & Snyder, 2008). All else equal, as measurement error increases, we expect responses to exhibit more variability across the same question.

In Study O, respondents were asked three questions twice, so we can measure response stability among respondents to surveys with and without attention screening. Each question asked respondents to estimate a quantity they might see reported in news media. In this case, respondents estimated the percentage of people using opioids illegally who have various traits, on a zero to 100 scale. Some respondents were randomly assigned to an informational treatment delivered between the two sets of estimation questions. The left panel of Figure 2 displays results from control groups in each wave, where we expect stable responses because no new information was provided. The mean absolute difference between the first and second estimates was lower for two of the three questions in the wave that only included participants who successfully passed the attention check. This finding suggests that, on average, responses were more consistent across repeated questions when using an attention screener.

The right panel of Figure 2 displays experimental results from each wave. The informational treatment contained an approximate answer to one of the questions (Question A), so we expect the treatment to reduce the absolute error in the second estimate of Question A, relative to the value provided in the informational treatment. In the first wave, with no attention screen, respondents exposed to the informational treatment were 5.0 (SE = 0.7) points more accurate than control. In the second wave, with an attention screen, treated respondents were 8.8 (SE = 0.5) points more accurate. By this measure, the strength of the manipulation was about 40% smaller when no attention screen was used (difference 3.8, SE = 0.9, $p < 0.001$). Given that reading and retaining factual information is a key component of many survey-based studies of digital media, inattentive respondents may lead to significant attenuation of treatment effect estimates.

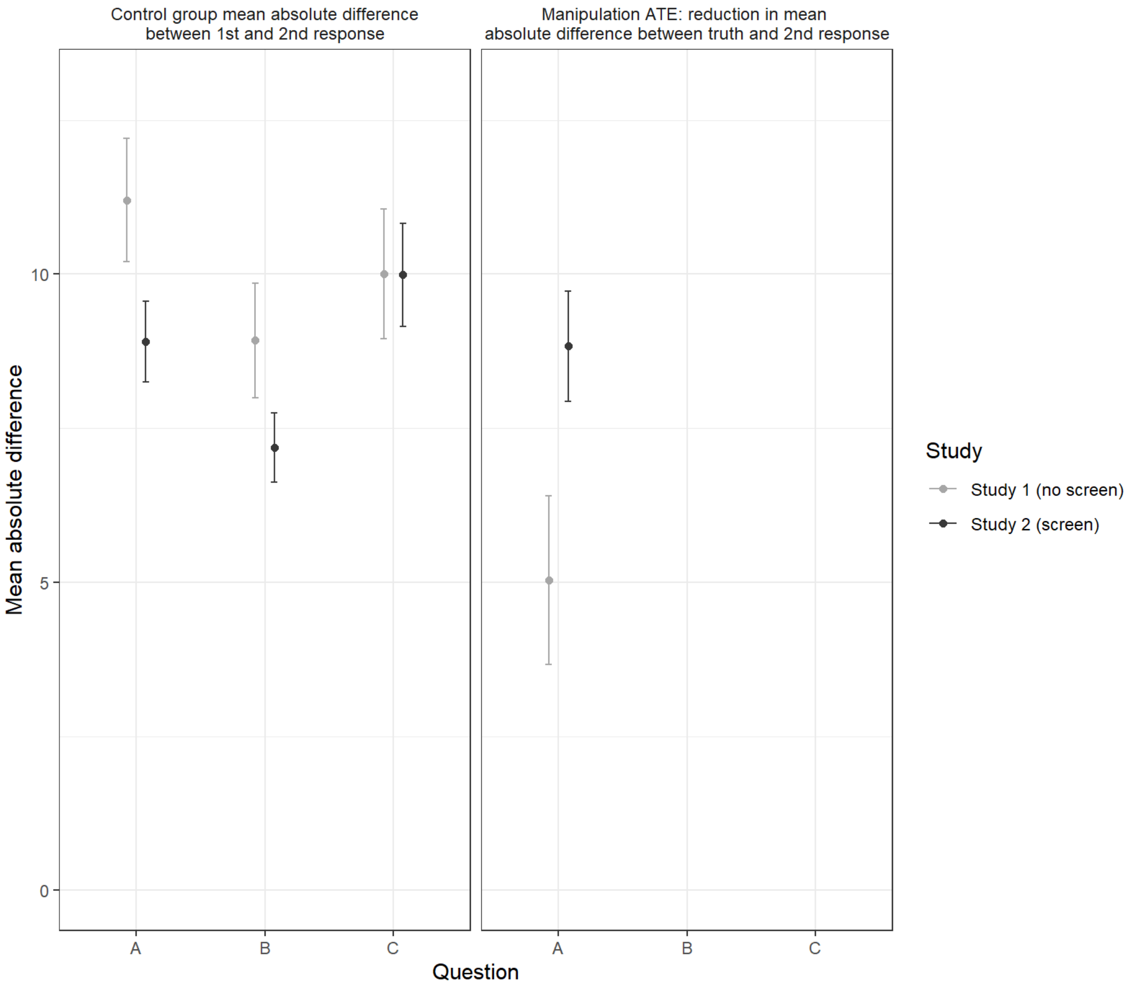

Figure 2. Survey Without Attention Screening had Lower Response Stability and Weaker Treatment Effects.

Note. All lines denote bootstrapped 95% confidence intervals. Mean absolute differences for Questions A and B are significant at $p < 0.01$. The difference in ATEs for Question A is significant at $p < 0.001$.

To illustrate the potential costs of failing to screen for attentiveness, we provide a simple power analysis in Figure 3. Specifically, we assess power to detect the manipulation ATE shown on the right of Figure 2. For sample sizes N ranging from 50 to 500, we repeatedly resampled N observations with replacement from the full set of consenting participants in each wave. We estimated the treatment effect using OLS in each sample to test the hypothesis that the effect was different from zero. In the survey wave with attention screening, we are overwhelmingly likely to detect a successful manipulation (at the 0.05 level) in samples of over 150 respondents, and we can achieve power of 0.8 with 66 respondents. In the wave without attention screening, we need approximately 250 respondents to detect a successful manipulation in 80% of simulated samples. As researchers move beyond detecting successful manipulation to detecting the effects of treatment on key outcomes, power is likely to remain substantially compromised if many respondents are not engaging with treatment materials.

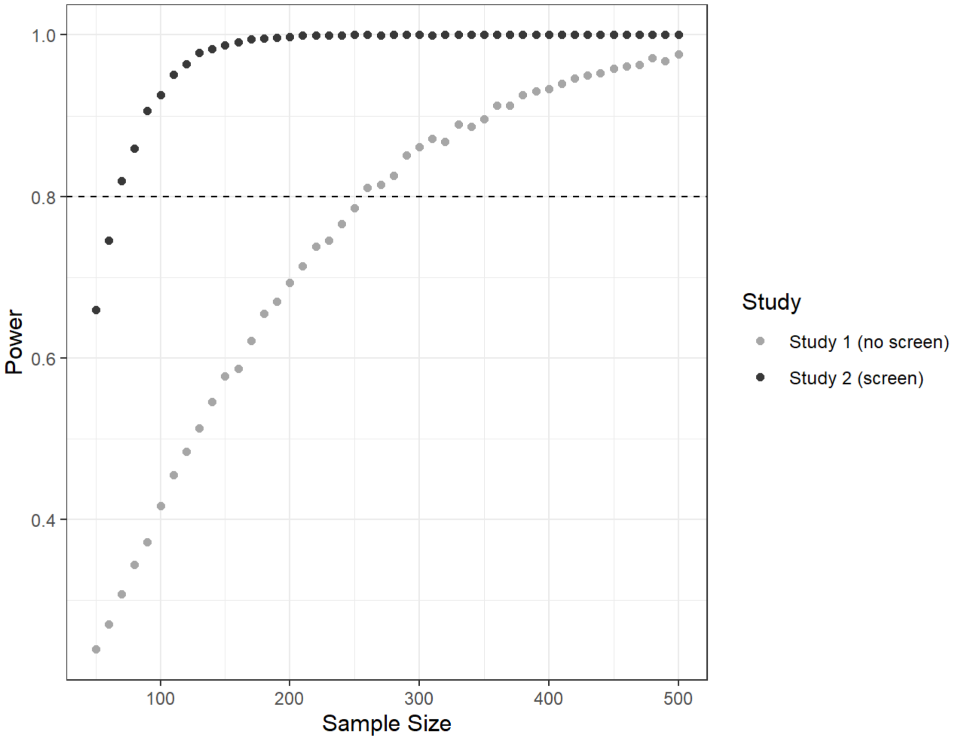

Figure 3. Survey Without Attention Screening Required Larger Sample Sizes to Detect a Treatment Manipulation Effect.

Note. Points indicate the percent of simulated studies of a given size in which the manipulation ATE (see Figure 2) was distinguishable from 0 using OLS with classical standard errors and a 0.05 significance threshold. Simulated studies drew from each wave of Study O, with replacement, drawing 5,000 samples with sizes that range from 50 to 500 in increments of 10. The dashed line indicates power of 0.8.

Demographics of Attentive and Inattentive Respondents

Consistent with prior research on attention checks (e.g., see Alvarez et al., 2019; Berinsky, Margolis, & Sances, 2014), we find that those who pass and those who fail attention checks are significantly different in terms of reported demographics (Table 1). Drawing on the demographic information Lucid collects from participants upon their enrollment in the platform,[17] participants who fail attention checks appear to be younger, more likely to be male, less likely to be college educated, and are more likely to have a household income below $15,000. Additionally, failing attention checks is associated with apparently lower levels of political knowledge. In one of the waves in Study T, participants were asked to identify Mitt Romney's political party before the attention checks. 85.0% of participants who passed all the attention checks were able to successfully identify Romney as Republican; only 30.2% of respondents who failed any one of the attention checks were able to correctly identify Romney's party (Pearson chi2(1) = 17.98, $p < 0.001$).

But these differences should be viewed with some level of skepticism, as we find that inattentive participants also have more inconsistencies when answering demographic questions. Namely, we attempted to confirm the identity of participants by comparing participants' responses to demographic questions in our survey with the demographic variables Lucid collected from participants upon their enrollment in the platform. Two of our studies asked for self-reported demographic information, S and T. In Study S, respondents were asked their state of residence. Among respondents who pass the attention checks, 96.5% have self-reported states that match Lucid location data. Among those who failed the attention checks, 80.7% have states that match (Pearson chi2(1) = 975.52, $p < 0.001$). In Study T, respondents were asked for their date of birth. Self-reported age in the survey was compared to the age variable collected by Lucid upon enrollment. Age matched 84.0% of the time among participants who passed all the attention checks and 75.6% of the time among those who failed at least one attention check (Pearson chi2(1) = 263.52, $p < 0.001$).

[17] Lucid participants are sourced from other suppliers and, in our experience taking Lucid surveys as participants, most demographic questions are asked (repeatedly) by different suppliers and not via Lucid-branded web pages. How that demographic information is compiled, validated, and eventually provided to the researcher is not specified by Lucid.

Table 1. Inattentive Respondents Differ Across Demographics

	Study K		Study T	
	Passed Attention Checks	Failed to Pass Attention Checks	Passed Attention Checks	Failed to Pass Attention Checks
Age	43.8	36.9	45.5	39.7
Female	59.3%	45.5%	52.6%	46.2%
HH Income Greater Than $100,000	15.0%	15.8%	21.2%	21.8%
HH Income Below $14,999	14.3%	26.9%	12.7%	22.0%
College Educated	38.7%	33.9%	49.4%	41.5%
Democrat	45.0%	44.5%	48.1%	44.8%

Note. Study K's LR chi2(6) = 5532.54 (p < 0.0001) and Study T's LR chi2(6) = 563.44 (p < 0.0001).

Trends in Measurement Error after Attention Screening

We have thus far demonstrated that attention screening may reduce measurement error, but our previous analyses do not rule out that measurement error may still be increasing over time even after screening for attentiveness. This latter possibility would imply that attention screens do not wholly ameliorate declines in data quality around the 2020 COVID-19 pandemic. To emphasize, attention screens are an imperfect means of excluding respondents who answer questions haphazardly (or without reading) and they do not address all forms of satisficing.[18]

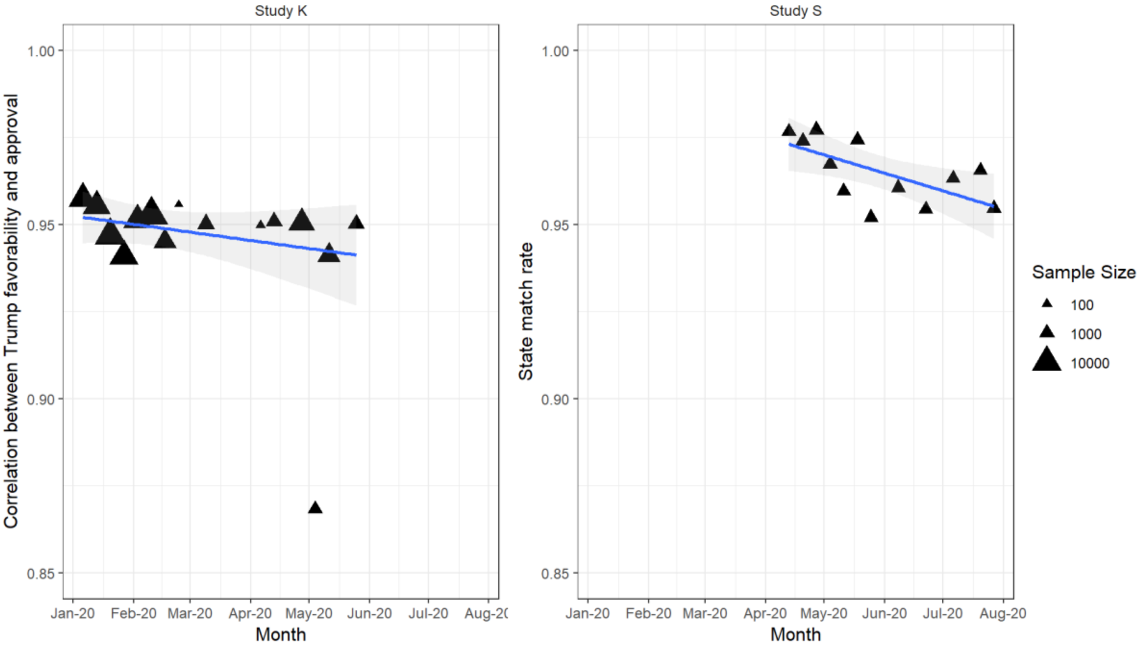

Figure 4. Mixed Evidence About Trends in Data Quality Among Attentive Respondents.

Note. In Study K (left panel), data quality is measured by correlating Trump approval and favorability. Study S (right panel) measured data quality as the match rate of self-reported state of residence and the state on file with Lucid.

We found mixed evidence as to whether data quality has decreased over time among participants who pass attention checks. As seen in Figure 4, Study S exhibited a slight but statistically significant decline in the rate at which respondents reported living in the same state indicated by Lucid ($p < 0.01$). Study K did not exhibit a statistically

[18] For instance, one can imagine a survey-taker who reads all questions carefully and thus passes all attention screens, but nevertheless selects nonsensical responses intentionally. As Abbey & Meloy (2017) note, attention checks may cause “annoyance regarding the researcher's intent (i.e., reactance).” (p. 65).

significant decline in the correlation between Trump approval and Trump favorability, both measured on seven-point scales (p = 0.19). These trends suggest that data quality among respondents who passed attention checks was not declining at the same rate of overall data quality in early 2020, but that screening on attention check questions is not sufficient to overcome all data quality challenges.[19]

Has Attentiveness Bounced Back in 2021?

After the first 6 months of 2020, further waves of Study T showed a slight increase in the rate at which Lucid respondents passed attention and audiovisual checks. During the initial waves from the first 6 months of 2020, Study T had a passage rate of 30.6% (N = 4,917). In additional waves from July, September, October and November 2020, Study T and had a passage rate of 42.9% (N = 7,362). Because this trend indicates the possibility of a rebound in attentiveness, we conclude by considering whether or not attention check passage rates have increased since the first 6 months of 2020.[20]

We draw on Study K to address this question because of its large size and continuation into 2021. Study K continued with additional waves in March, April and June 2021. In the first three months of 2020, the attention check passage rate in study K was 83.2% on average (N = 80,237). In the second three months, it fell to 78.3% (N = 19,367). Across all of the additional waves from 2021, the passage rate for identical attention check measures was 77.1% (N = 19,568). This represents a further decrease of 1.1% since the waves launched in April through June 2020 (p < 0.01). Study K thus provides no evidence of a rebound in attentiveness in early 2021.

[19] Ideally, we would want to examine the same trend for those participants who fail attention checks, however, this data is not available; both these studies terminated the surveys of inattentive participants before they could complete the entire survey.

[20] We note as well that Lucid has been aware of the decline in attention check passage rates and may have changed internal quality control measures. No summary of these product changes is publicly available. However, a Lucid representative confirmed that the company does have procedures to detect bots and has an internal library of attention check questions in addition to a Quality Program (https://luc.id/quality/), which formally assesses quality for each supplier that Lucid uses for survey distribution.

Discussion and Conclusions

Although a large literature exists assessing the overall quality of various online survey platforms, this paper illustrates that platforms' data quality should be periodically reassessed. Arechar and Rand (2021) found similar declines in attentiveness on Amazon's Mechanical Turk around the time of the COVID-19 pandemic. However, online survey data quality isn't always dependent on catastrophic external events such as a global pandemic. Dennis, Goodson, and Pearson (2020) found that even before the pandemic, there were data quality issues on Mturk due to the prevalence of services that mask an individual survey-taker's identity. More recently, Prolific experienced a massive influx of young, female survey-takers when the Prolific platform unexpectedly went viral on the popular social media website, TikTok; this led to surveys with less than 25% male participants (Charalambides 2021). This is all to say that even reputable, previously-validated online survey platforms may experience unexpected shocks in data quality. Even in the absence of well-documented external events, data quality may fluctuate over time.

On Lucid, we found meaningfully large differences in attentive respondents over time. The same survey and the same attention check could yield an attention check success rate of 85% in one week and yield a success rate that is 15 percentage points lower in a different week. We find that inattentive survey-takers give less consistent responses and less reliable demographic information. Most crucially, we find that failing to screen for inattentive survey-takers can meaningfully reduce the size of treatment effects in online survey experiments. In other words, if a researcher fails to screen on attentiveness, an otherwise well-powered experiment could nevertheless yield false negative results despite there being a true underlying treatment effect. Our findings predict widespread attenuation of treatment effect estimates during our study period. Such attenuation is observed in Peyton, Huber, and Coppock's (2021) replication of survey experiments on Lucid.

While attention checks are a popular way of detecting and excluding participants with potentially high levels of measurement error, attention checks do contain drawbacks and are not a panacea. For example, screening for respondents who pass attention checks may lead to artificially large treatment effects if these respondents are more likely to engage with a textual treatment, or artificially small treatments effects if these respondents are already more knowledgeable and therefore potentially less persuadable (compared to a representative sample measured without measurement error). Our findings illustrate that this trade-off between measurement error and sample composition must be carefully considered by researchers using Lucid, or likely any online convenience sample. Moving forward, we recommend that researchers should clearly report if their survey incorporated attention or other data quality checks. If they have not, these researchers may wish to consider replicating any null findings, as they may have been driven by attenuation bias.

Furthermore, due to a slight decline in location match rate even among those who successfully pass the attention check in one study, we recommend that researchers incorporate other data quality checks in all online surveys to guard against sudden declines in data quality. One way to do this would be to look at the consistency of responses across highly similar questions. As such, when it is possible, researchers should consider including variants of the same question throughout their survey to assess the extent of possible measurement error. We also recommend asking participants their date of birth and/or location to match to the demographic data provided by the online platform. While a mismatch does not necessarily mean there is measurement error, we've documented a clear association between demographic response mismatch and inattentiveness.

This paper used several metrics to estimate data quality (i.e., consistency of responses, reliability of demographic information, ability to answer direct query attention checks), but no one metric should be relied on exclusively. In other words, poor scores on these metrics can be viewed as symptoms of bad data. As is the case with symptoms of diseases, some subset of these symptoms will be false positives, but when taken together they paint a far more compelling picture of the underlying quality of the data. Researchers should take steps to confirm that each respondent corresponds to a single, independent participant; that participants are who they say they are; and participants aren't selecting responses without reading the survey question. These are difficult phenomena to diagnose, but a failure to do so may lead to false negative null effects.

Acknowledgements

We are immensely grateful to Joshua Kalla and P. M. Aronow for their invaluable contributions to this project. We also thank Brian Schaffner for sharing his data to be included in this analysis.

References

Abbey, J. D., & Meloy, M. G. (2017). Attention by design: Using attention checks to detect inattentive respondents and improve data quality. *Journal of Operations Management, 53*, 63-70.

AJMC Staff. (2021). A Timeline of COVID-19 Developments in 2020. *American Journal of Managed Care.* https://www.ajmc.com/view/a-timeline-of-covid19-developments-in-2020

Alvarez, R. M., Atkeson, L. R., Levin, I., & Li, Y. (2019). Paying attention to inattentive survey respondents. *Political Analysis, 27*(2), 145-162.

Amir, O., Rand, D. G., & Gal, Y. A. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS one*, 7(2), e31461.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, 45(6), 842-850.

Ansolabehere, S., Rodden, J., & Snyder, J. M. (2008). The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting. *American Political Science Review* 102: 215–32.

Arechar, A. A., & Rand, D. G. (2021). Turking in the time of COVID. *Behavior research methods*, 1-5.

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*(4), 572-589.

Atkeson, L. R., & Alvarez, R. M. (Eds.). (2018). *The Oxford handbook of polling and survey methods*. Oxford University Press.

Benzell, S. G., Collis, A., & Nicolaides, C. (2020). Rationing social contact during the COVID-19 pandemic: Transmission risk and social benefits of US locations. *Proceedings of the National Academy of Sciences*.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political analysis*, 20(3), 351-368.

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2014). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science*, *58*(3), 739-753.

Berinsky, A. J., Margolis, M. F., Sances, M. W., & Warshaw, C. (2021). Using screeners to measure respondent attention on self-administered surveys: Which items and how many?. *Political Science Research and Methods*, 9(2), 430-437.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?. *Perspectives on Psychological Science*, 3-5.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, 13(2), 149-154.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods*, 46(1), 112-130.

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. A. (2015). Using nonnaive participants can reduce effect sizes. *Psychological science*, 26(7), 1131-1139.

Charalambides, N. (2019). We recently went viral on TikTok - here's what we learned. *Prolific*. August 24, 2021. https://blog.prolific.co/we-recently-went-viral-on-tiktok-heres-what-we-learned/

Conn, K. M., Mo, C. H., & Sellers, L. M. (2019). When less is more in boosting survey response rates. *Social Science Quarterly*, *100*(4), 1445-1458.

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, 7(3), 613-628.

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, 115(49), 12441-12446.

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research & Politics*, *6*(1), 2053168018822174.

Dennis, S. A., Goodson, B. M., & Pearson, C. A. (2020). Online worker fraud and evolving threats to the integrity of MTurk data: A discussion of virtual private servers and the limitations of IP-based screening procedures. *Behavioral Research in Accounting*, 32(1), 119-134.

Downs, J. S., Holbrook, M. B., & Peel, E. (2012). Screening participants on Mechanical Turk: Techniques and justifications. *ACR North American Advances*.

Goodman, J. K., & Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research*, 44(1), 196-210.

Groves, R. M. (1987). Research on survey data quality. *The Public Opinion Quarterly*, *51*, S156-S172.

Gummer, T., Roßmann, J., & Silber, H. (2018). Using instructed response items as attention checks in web surveys: Properties and implementation. *Sociological Methods & Research*, 0049124118769083.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary?. *Frontiers in psychology*, *9*, 998.

Hauser, D. J., & Schwarz, N. (2015). It'sa trap! Instructional manipulation checks prompt systematic thinking on “tricky” tasks. *Sage Open*, *5*(2), 2158244015584617.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, *48*(1), 400-407.

Hemel, D., & Porter, E. (2016). Aligning taxes and spending: theory and experimental evidence. *Behavioural Public Policy*, 1-21.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. (2020). The shape of and solutions to the MTurk quality crisis. Political Science Research and Methods, 8(4), 614-629.

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public opinion quarterly*, 51(2), 201-219.

Kung, F. Y., Kwok, N., & Brown, D. J. (2018). Are attention check questions a threat to scale validity?. *Applied Psychology*, *67*(2), 264-283.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of Mechanical Turk samples. *Sage Open*, 6(1), 2158244016636433.

Lucid. (2021, July 26). Academic Studies. Retrieved from https://luc.id/citations/

Malhotra, N. (2008). Completion time and response order effects in web surveys. *Public opinion quarterly*, 72(5), 914-934.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological methods*, *17*(3), 437.

Moreland, A., Herlihy, C., Tynan, M. A., Sunshine, G., McCord, R. F., Hilton, C., ... & Popoola, A. (2020). Timing of state and territorial COVID-19 stay-at-home orders and changes in population movement—United States, March 1–May 31, 2020. *Morbidity and Mortality Weekly Report*, *69*(35), 1198.

Mortensen, K., & Hughes, T. L. (2018). Comparing Amazon's Mechanical Turk platform to conventional data collection methods in the health and medical research literature. *Journal of General Internal Medicine*, 33(4), 533-538.

Motta, M., Chapman, D., Stecula, D., & Haglin, K. (2019). An experimental examination of measurement disparities in public climate change beliefs. *Climatic change*, *154*(1-2), 37-47.

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science*, 2(2), 109-138.

Peyton, K., Huber, G. A., & Coppock, A. (2021). The Generalizability of Online Experiments Conducted During the COVID-19 Pandemic. *Journal of Experimental Political Science*, 1–16.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science*, *31*(7), 770-780.

Paas, L. J., & Morren, M. (2018). Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters*, *29*(1), 13-21.

Read, B., Wolters, L., & Berinsky, A. J. (2020). Racing the Clock: Using Response Time as a Proxy for Attentiveness on Self-Administered Surveys.

Roberts, C., Gilbert, E., Allum, N., & Eisner, L. (2019). Research synthesis: Satisficing in surveys: A systematic review of the literature. *Public Opinion Quarterly*, 83(3), 598-626.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of personality and social psychology*, 51(3), 515.

Silber, H., Danner, D., & Rammstedt, B. (2019). The impact of respondent attentiveness on reliability and validity. *International Journal of Social Research Methodology*, *22*(2), 153-164.

Solnick, R. E., Peyton, K., Kraft-Todd, G., & Safdar, B. (2020). Effect of Physician Gender and Race on Simulated Patients' Ratings and Confidence in Their Physicians: A Randomized Trial. *JAMA network open*, *3*(2), e1920511-e1920511.

Stewart, N., Ungemach, C., Harris, A. J., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision making*, 10(5), 479-491.

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing samples in cognitive science. Trends in cognitive sciences, 21(10), 736-748.

Wood, D., Harms, P. D., Lowman, G. H., & DeSimone, J. A. (2017). Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science*, *8*(4), 454-464.

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135-163.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of personality and social psychology*, 111(4), 493.

Appendix

Table OA1: Papers that Use Lucid Data Published in Top Political Science Journals in the Past Three Years

Citation	Journal
Orr, L. V., & Huber, G. A. (2020). The policy basis of measured partisan animosity in the United States. *American Journal of Political Science*, *64*(3), 569-586.	AJPS
Klar, S., & McCoy, A. (2021). Partisan-Motivated Evaluations of Sexual Misconduct and the Mitigating Role of the# MeToo Movement. *American Journal of Political Science*, *65*(4), 777-789.	AJPS
Offer-Westort, M., Coppock, A., & Green, D. P. (2021). Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*.	AJPS
Costa, M. (2021). Ideology, not affect: What Americans want from political representation. *American Journal of Political Science*, *65*(2), 342-358.	AJPS
Hill, S. J., & Huber, G. A. (2019). On the Meaning of Survey Reports of Roll-Call "Votes". *American Journal of Political Science*, *63*(3), 611-625.	AJPS
Guay, B., & Johnston, C. D. (2020). Ideological asymmetries and the determinants of politically motivated reasoning. *American Journal of Political Science*.	AJPS
Peterson, E., & Kagalwala, A. (2021). When unfamiliarity breeds contempt: how partisan selective exposure sustains oppositional media hostility. *American Political Science Review*, *115*(2), 585-598.	APSR
Tomz, M., & Weeks, J. L. (2020). Public opinion and foreign electoral intervention. *American Political Science Review*, *114*(3), 856-873.	APSR
Graham, M. H., & Svolik, M. W. (2020). Democracy in America? Partisanship, polarization, and the robustness of support for democracy in the United States. *American Political Science Review*, *114*(2), 392-409.	APSR
Myrick, R. (2020). Why So Secretive? Unpacking Public Attitudes toward Secrecy and Success in US Foreign Policy. *The Journal of Politics*, *82*(3), 828-843.	JOP

Coppock, A., & Green, D. P. (2017). Do Belief Systems Exhibit Dynamic Constraint?. *Journal of Politics*.	JOP
Schwarz, S., & Coppock, A. (2020). What have we learned about gender from candidate choice experiments? A meta-analysis of 67 factorial survey experiments.	JOP
Kennedy, R., Waggoner, P., & Ward, M. (2018). Trust in public policy algorithms. *Journal of Politics*.	JOP
Levy, M. E. (2021). Once Racialized, Now "Immigrationized"? Explaining the Immigration-Welfare Link in American Public Opinion. *The Journal of Politics, 83*(4), 1275-1291.	JOP
Tomz, M. R., & Weeks, J. L. (2020). Human rights and public support for war. *The Journal of Politics*, *82*(1), 182-194.	JOP
Lajevardi, N. (2021). The media matters: Muslim-American portrayals and the effects on mass attitudes. *The Journal of Politics*, *83*(3).	JOP

Table OA2: Regression models of time trend in attention check passage

		Passage Rate	
	Study K	Study S	Study T
Intercept (Jan 1, 2020)	0.851 [0.831, 0.872]	0.976 [0.522, 1.43]	0.839 [0.212, 1.466]
Week	-0.004 [-0.007, -0.001]	-0.012 [-0.03, 0.014]	-0.028 [-0.061, 0.005]
N Weeks	16	9	3
N Responses	99,600	10,889	4,917

Note: This table presents OLS coefficients and 95% confidence intervals calculated with robust standard errors for the models underlying Figure 1.

Table OA3: Robustness of time trend in attention check passage

	Passage Rate	
	First 6 Months of 2020	2020
Intercept (Jan 1, 2020)	0.853 [0.833, 0.872]	0.831 [0.797, 0.865]
Week	-0.005 [-0.007, -0.002]	-0.001 [-0.006, 0.003]
Study S	-0.014 [-0.083, 0.054]	-0.056 [-0.136, 0.025]
Study T	-0.462 [-0.509, -0.414]	-0.409 [-0.625, -0.194]
N Weeks	28	35
N Responses	115,406	126,182

Note: This table presents OLS coefficients and 95% confidence intervals calculated with robust standard errors for additional models of the time trend in attention/ audiovisual check passage beyond what is reported in the main.